Postprint version of paper published in Process Biochemistry 2013, vol. 48(11), p. 1789–1793.
The content is identical to the published paper, but without the final typesetting by the publisher.

Multivariate assessment of activated sludge stability in lab-scale experiments

P. Van den Kerkhof[a], G. Gins[a], R. Van den Broeck[a], J.F.M. Van Impe[a,*]

[a] *KU Leuven, Department of Chemical Engineering, Chemical & Biochemical Process Technology & Control (BioTeC), W. de Croylaan 46, B-3001 Leuven, Belgium*

Abstract

In wastewater research, the transfer of activated sludge from large-scale plants to lab-scale reactors induces a transient period. It is crucial to quantitatively assess the stability of activated sludge before starting any experimental procedure. Otherwise, the transient interferes with the experimental results, jeopardizing reproducibility and accuracy. This paper presents a novel *multivariate* technique to assess activated sludge stability based on the combination of principal component analysis (PCA) with the squared prediction error (SPE) statistic. The proposed method allows for a more accurate estimation of activated sludge stability than existing *univariate* methods and also eliminates the need to establish thresholds for every single variable. The procedure is validated on experimental data obtained by Van den Broeck et al. [1].

Keywords: Biological wastewater treatment, Activated sludge stability, Transient, Principal component analysis (PCA), Fault detection

1. Introduction

Most large-scale wastewater treatment plants (WWTPs) employ the activated sludge process. For practical reasons, research on activated sludge processes is often performed on a lab-scale. Sludge needed to inoculate lab-scale experiments is typically taken from large WWTPs to ensure a diverse and representative microbial community. However, the transfer to lab-scale reactors induces a transient period during which the micro-organisms adapt

*Corresponding author: jan.vanimpe@cit.kuleuven.be; tel. +32 16321466

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

to the new environment [2]. It is crucial to quantitatively assess whether the activated sludge has reached stable conditions before starting the experimental procedure to avoid biased experimental results.

Traditionally, the sludge age is used as stability index by allowing $2-3\times$ the sludge age as adaptation period. As argued in Van den Broeck et al. [1], this period can be either too short (e.g., if an unforeseen disturbance took place) or too long (delaying the start-up of the experiment). Barbusiński and Miksch [3] and Higgins and Novak [4] assessed stability by requiring the difference between subsequent measurements on some selected variables to be smaller than $10-20\%$ for a specified duration. Van den Broeck et al. [1] proposed a 7-day moving window approach, during which the mean slope and the gap between maximum and minimum of the measurement profiles need to be smaller than a threshold based on a reference experiment.

The above approaches are essentially *univariate*, i.e., each selected variable is treated as a separate stability index for which thresholds need to be defined. Moreover, not every available variable is suited for *univariate* stability assessments. For example, Van den Broeck et al. [1] disposed of multiple morphological variables which exhibited a stabilization trend, but were numerically ineffective as stability indicators due to measurement noise.

This paper proposes a novel *multivariate* stability assessment technique based on principal component analysis (PCA) [5, 6], a technique previously applied successfully to wastewater applications (see e.g., [7, 8, 9, 10]). The proposed approach only requires a rough variable selection by removing variables which are known to be unrelated to sludge stability and retaining those variables that have a possible impact on stability. The stability information in the different variables is summarized in one scalar stability index, i.e., the squared prediction error statistic. This way, only one overall threshold needs to be defined instead of establishing thresholds for every selected variable. Furthermore, taking into account the variable correlations in PCA increases the ability to detect disturbances which prolong the transient period.

2. Materials and methods

2.1. Stability assessment procedure

Step 1: Reference experiments. The first step consists of conducting reference experiments to obtain training data for the PCA model. The experimental set-up is inoculated and monitored until the measurements are visually stable, i.e., displaying only small fluctuations for an extended period of time

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

(e.g., 7 days). The duration of the reference experiments should be long enough to reach stability and to collect a sufficient amount of training data characterizing stable behavior.

Step 2: Training set definition. During the second step, the data of each reference experiment is divided into two parts: an initial (potentially) unstable part and a final (certainly) stable part. Only the latter part will be used as training data. It is crucial to exclude data reflecting unstable operation from the training set, otherwise the fault detection statistic's ability to discriminate between stable and unstable operation will be seriously impaired. Therefore, only the last few data points of each reference experiment, where stability is almost certain, are included in the training set at this step.

Step 3: Model identification. In step 3, the principal components and the fault detection statistic's control limit are identified on the training data. The number of principal components is determined with a criterion similar to the adjusted Wold R criterion for partial least squares [11]. The first r components are included until

$$\frac{\sum_{i=1}^{r+1} f_i}{\sum_{i=1}^{r} f_i} < 1.05 \tag{1}$$

where f_i is the variance explained by the i-th principal component. Of the two most popular fault detection statistics, the Hotelling's T^2 statistic and the squared prediction error (SPE), only the latter is employed in this work. This choice is justified in the Section 2.2. The SPE equals the squared sum over all J variables of the difference between the actual value x_j and PCA model fit $\hat{x}_j$ and has an upper control limit (UCL).

$$\mathrm{SPE} = \sum_{j=1}^{J} (x_j - \hat{x}_j)^2 \tag{2}$$

$$\mathrm{UCL_{SPE}} = \frac{\sigma^2_{\mathrm{SPE}}}{2\mu_{\mathrm{SPE}}}\chi^2(\frac{2\mu^2_{\mathrm{SPE}}}{\sigma^2_{\mathrm{SPE}}};\alpha) \tag{3}$$

The mean and variance of the SPE statistic over all training sample are denoted by μ_{SPE} and σ^2_{SPE} and $\chi^2(2\mu^2_{\mathrm{SPE}}/\sigma^2_{\mathrm{SPE}};\alpha)$ is the upper critical value of a χ^2-distribution with $2\mu^2_{\mathrm{SPE}}/\sigma^2_{\mathrm{SPE}}$ degrees of freedom at a specified tolerance level α [12]. The model is now ready to monitor unseen data points.

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

Postprint version of paper published in Process Biochemistry 2013, vol. 48(11), p. 1789–1793.
The content is identical to the published paper, but without the final typesetting by the publisher.

Step 4: Control chart generation. During step 4, the presumed unstable data is added again and, using the PCA model from the previous step, the fault detection statistic for the entire reference data is computed. The value of the statistic versus time and its control limit are plotted on a control chart.

Step 5: Control chart evaluation. The training set is initially kept small to ensure that data from unstable operation is excluded from the training set, as it would degrade the control chart's performance. However, increasing the size of the training set increases the PCA model's robustness and reliability. Therefore, the initial division between an unstable and a stable region is reviewed in step 5 by studying the control charts.

Each control chart exhibits a trend of values exceeding the UCL at the beginning of the experiment before decreasing below the control limit. If the initial stability region is smaller than the region determined from the control chart, the training set is expanded and the PCA model retrained. At this step it is important to keep in mind that a point exceeding the UCL is proof of instability, whereas a value of the statistic below the UCL is only the *absence of proof* for instability since the null hypothesis can not be rejected. Hence, it is not recommended to expand the presupposed stability region until all data points having a statistic below the UCL are included. It is advisable to exclude a small data window (i.e., stability horizon) starting from the point where the statistic is below its UCL to ensure that the activated sludge is in a stable state. The dynamics of the transient period, i.e., the steepness of the descent crossing the control limit, determine the appropriate window length. Steps 3 to 5 are repeated until the presumed stability region and the region determined from the control charts agree.

Step 6: Validation. After the training procedure, the identified PCA model is used to generate control charts to monitor the start-up of new experiments. As soon as the statistic remains below its control limit for a number of consecutive days, i.e., the window length determined in the training procedure, the activated sludge is considered stable and the intended experimental procedure can be started without transients interfering with the results.

2.2. Data scarcity

The amount of training data for activated sludge stability assessments is severely limited compared to typical PCA based monitoring applications. This presents a challenge for the determination of the UCLs. Ramaker et al.

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

[13] reported that for the SPE, the false alarm rate is too high compared to the specified significance level α, i.e., the computed UCL is too low. Conversely, the UCL of the T^2 is too high. This effect is small for large data sets, but becomes very pronounced for small training sets as in this application.

Ramaker et al. [13] proposed a leave-one-out (LOO) procedure to alleviate this problem. Each training point is left out of the training set once and the principal components are computed from the remaining data. Subsequently, the LOO scores and residuals of the left-out data point are computed. The LOO scores and residuals are more comparable to validation scores and residuals. Ramaker et al. [13] reported that the LOO procedure significantly reduces the problem of too high false alarm rates for the SPE but the T^2 statistic's UCL remains troublesome. Therefore, only the SPE is employed in this work. The mean μ_{SPE} and variance σ_{SPE} in the SPE's UCL (see Eq. 3) are computed by calculating the SPE of each LOO residual using Eq. 2 and computing its mean and variance, respectively.

2.3. Case study

The proposed stability assessment procedure is validated on the experimental data generated by Van den Broeck et al. [1]. The experimental set-up consisted of a cylindrical bioreactor with a volume of 20 liter connected to a conic settler with a volume of 5 liter. The bioreactor was inoculated with activated sludge from a municipal wastewater treatment facility (Leuven, 130 000 population equivalent) and fed with synthetic influent. The sludge age was 20 days. Three experiments were conducted during which daily measurements were taken for a period of 40 days. Both conventional properties (e.g, SVI, MLSS) and morphological properties, acquired via microscopic image analysis, were measured. The sludge properties included in the dataset for this study are summarized in Table 1. The reader is referred to Van den Broeck et al. [1] for a detailed description of the experimental set-up.

3. Results and discussion

3.1. Reference experiment

Fig. 1 depicts the measurement profiles of a few selected variables from those recorded during the reference experiment. All profiles exhibit a decreasing or increasing trend during the adaptation phase however, not all of them are suited as a univariate stability index. The profiles of the SVI and filament length both settle to a nearly constant value whereas the remaining

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

Table 1: Selected variables from the experimental data set of Van den Broeck et al. [1].

Name	Abbreviation	Range
Biochemical variables		
Maximum specific oxygen uptake rate	$SOUR_{max}$	10–125 $mgO_2/gMLSS{\cdot}h$
Physical variables		
Sludge volume index	SVI	45–225 mL/g
Mixed liquor suspended solids	MLSS	2.25–5.65 g/L
Morphological variables		
Number of flocs	Nv	12–40
Number of filaments	Nf	17–160
Number of filaments, flocs and fragments	Nt	180–1300
Aspect ratio	AR	1.85–2.25
Convexity	C	0.60–0.75
Compactness	CP	0.65–0.75
Fractal dimension	FD	1.20–1.35
Form factor	FF	0.25–0.40
Roundness	R	0.45–0.55
Reduced radius of gyration	RG	0.80–0.95
Solidity	S	0.70–0.80
Equivalent floc diameter	D_{eq}	40–95 pixels
Filament length	F	400–7400 pixels
Total floc area	Av	1200–7100 $pixels^2$
Total area of flocs and filaments	Avf	30,000–95,000 $pixels^2$

variables still exhibit larger fluctuations. High noise levels render criteria on the maximal gap or derivative during a predefined window ineffective. For example, Van den Broeck et al. [1] selected only the SVI, $SOUR_{max}$ and filament length out of the available variables in Table 1. In the proposed multivariate approach all available variables are included. As a consequence, the trend information contained in noisy signals is preserved and the ability to detect disturbances possibly prolonging the transient phase improves.

Only the last 5 measurement days are retained to define the initial stability region. A PCA model is identified according to the procedure of Section 2.1; the adjusted Wold R criterion yields 5 principal components. Using the PCA model, the full reference data is monitored resulting in the control chart depicted in Fig. 2(a). The UCL has a significance level α of 0.1%.

The control chart exhibits a clear distinction between an unstable part with high values of the SPE and a stable part in which the SPE is below its UCL. Zooming in on the control limit (Fig. 2(b)) reveals that the SPE is below its UCL for 8 subsequent measurements before reaching the presup-

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

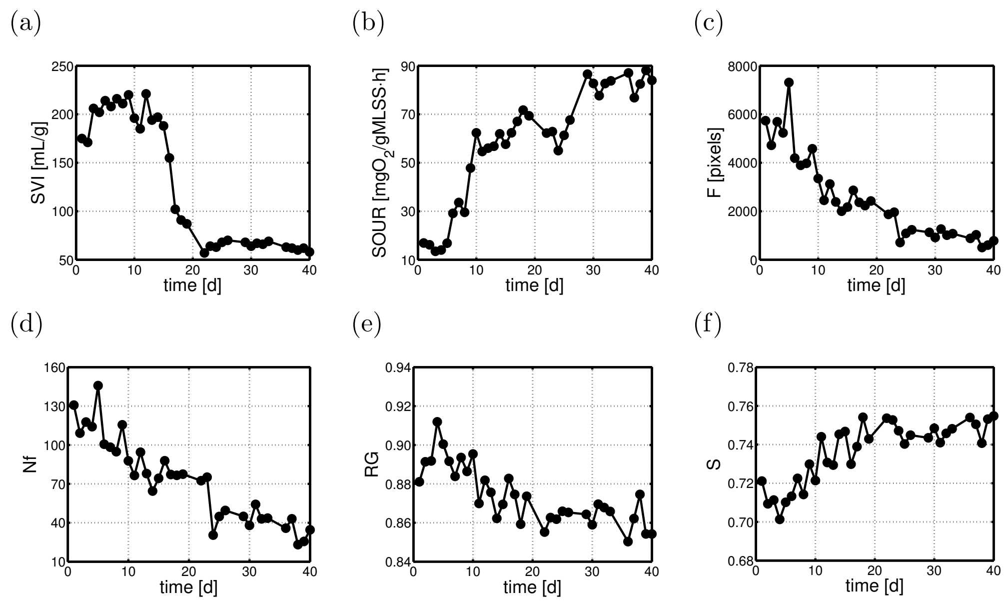

Figure 1: Reference experiment measurement profiles of the (a) sludge volume index SVI, (b) maximum specific oxygen uptake rate $\mathrm{SOUR}_{\mathrm{max}}$, (c) filament length F, (d) number of filaments Nf, (e) reduced radius of gyration RG and (f) solidity S.

posed initial stability region. Hence, expanding the training set is justified. However, as explained in Section 2.1, it is not recommended to expand the training set with all data points having a statistic below the UCL. In this case study, a stability horizon of 5 measurement days is chosen. Hence, the stability region is expanded with 3 measurement days and the PCA model retrained. The steps of retraining the PCA model and expanding the training set are repeated until the size of the training set and the stability region determined from the control chart agree. Figs. 2(c) and (d) depicts the resulting final control chart of the reference experiment. The final PCA model is trained on the last 11 data points and contains 5 principal components.

3.2. Validation experiments

The PCA model and control limits are validated on two experiments. The control chart of the first validation experiment is depicted in Fig. 3(a) and (b). Respecting a data window of 5 points, the sludge is stable from day 25 on. However, Van den Broeck et al. [1] reported the stable period to begin at day 37. In the *univariate* approach of Van den Broeck et al. [1], each selected

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

Postprint version of paper published in Process Biochemistry 2013, vol. 48(11), p. 1789–1793.
The content is identical to the published paper, but without the final typesetting by the publisher.

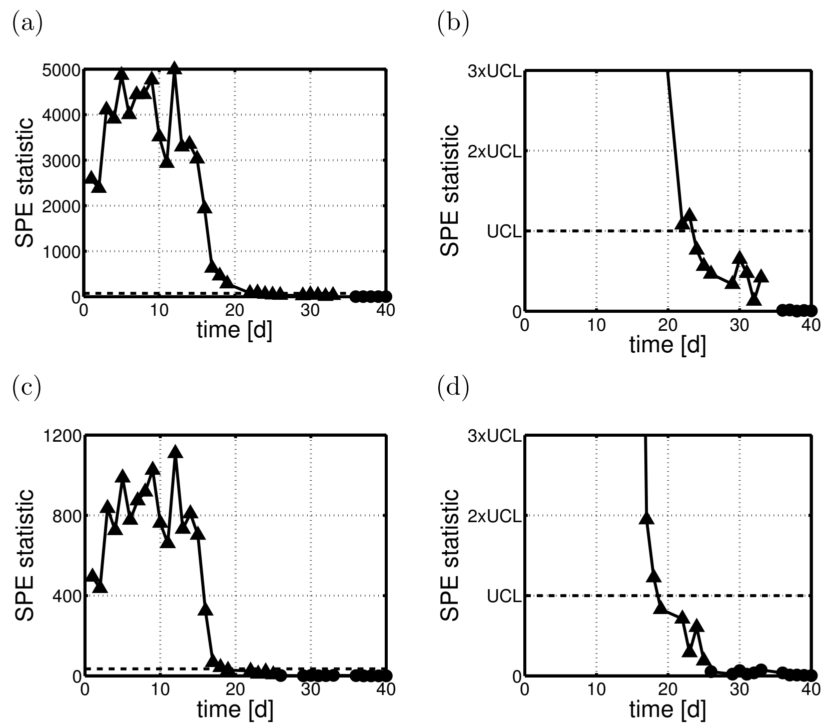

Figure 2: Illustration of the training procedure: (a) the initial control chart, (b) the initial control chart zooming in on the control limit, (c) the final control chart and (d) the final control chart zooming in on the control limit. A • represents a data point used to train the PCA model and a ▲ indicates validation data.

variable satisfies its own stability criterion before the sludge is considered stable. Although the filament length reaches stable conditions from day 24 on, fluctuations in the SVI and $SOUR_{max}$ prevent the *univariate* stability of each selected variable until day 37. According to the identified PCA model, these fluctuations do not significantly differ from the fluctuations during the reference experiment's stable period. As a result, stability is detected earlier in the proposed *multivariate* approach.

The control chart of the second validation experiment is depicted in Fig. 3(c) and (d). Due to technical problems, no $SOUR_{max}$ measurements are available after day 15. Missing measurements are a common occurrence in activated sludge experiments, which typically span over several weeks. Arteaga and Ferrer [14] provided an overview of methods for treating missing data in PCA. The simple trimmed score method is adopted in this work. Again tak-

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

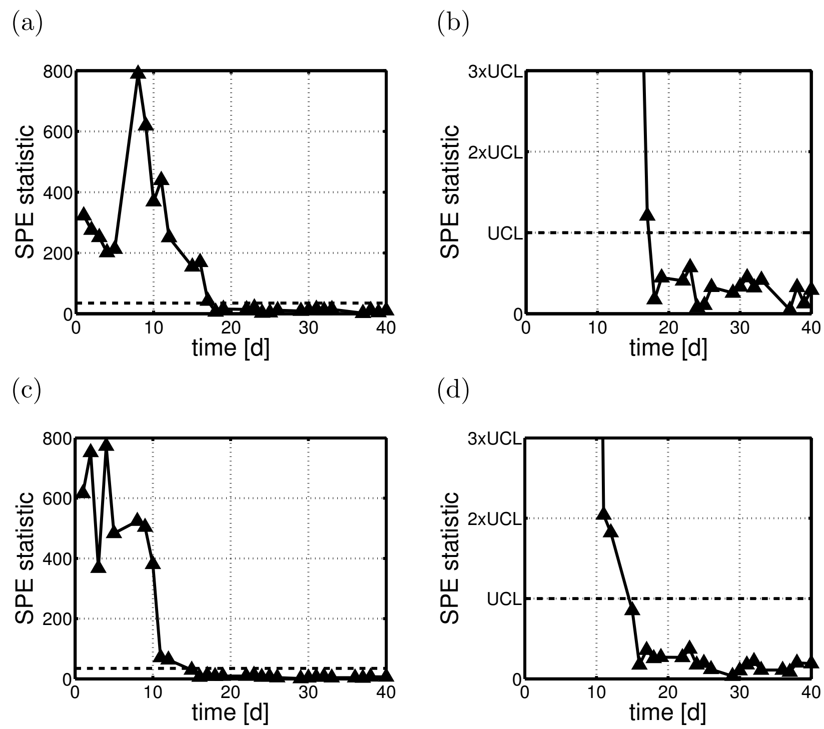

Figure 3: Control charts of the validation experiments: (a) Control chart of the first experiment, (b) the same control chart zooming in on the control limit, (c) control chart of the second experiment and (d) the same control chart zooming in on the control limit.

ing into account a 5 day window, the sludge is stable from day 22 on, which is in close agreement with Van den Broeck et al. [1] who reported a stable period beginning at day 21 after discarding the $SOUR_{max}$ measurements.

3.3. Post experiment variable selection

In the presented case study, every available variable associated with sludge characteristics or sludge stability is included in the PCA model. Hence, to compute the control charts all the variables included in the model need to be measured. To determine whether the experimental burden can be reduced for future experiments, the assessment procedure can be repeated after removing one or more variables. If the control charts still discriminate between unstable and stable, the removed variables can be omitted.

Figs. 4(a)-(d) depicts the control chart of the reference experiment after leaving out the SVI and $SOUR_{max}$, respectively. The two partial PCA models

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

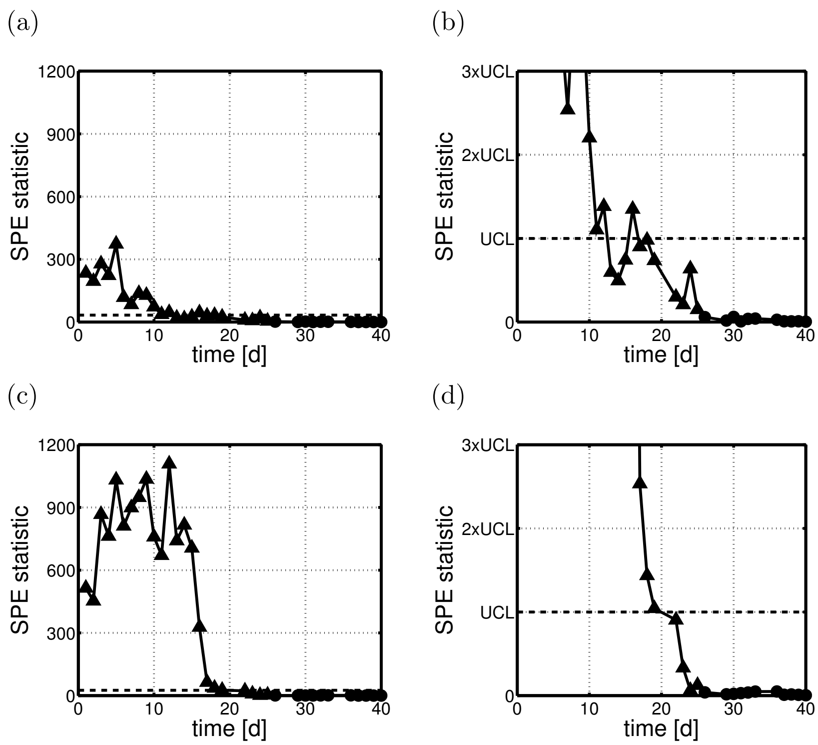

Figure 4: Illustration of post experiment variable selection: (a) control chart leaving out the SVI, (b) the same control chart zooming in on the control limit, (c) control chart leaving out the $SOUR_{max}$ and (d) the same control chart zooming in on the control limit. A • represents a data point used to train the PCA model and a ▲ indicates validation data.

are trained on the same amount of training points as the full PCA model. In Figs. 4(a) and (b), the difference between unstable and stable severely diminishes. Hence, the SVI is an informative measurement and should be retained. The $SOUR_{max}$ however, can be omitted as Figs. 4(c) and (d) still exhibit good discrimination between the transient and stable period. This means that, in this experimental set-up, the other measured variables contain enough information for a PCA model to still adequately discriminate between transient and stable sludge.

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

4. Conclusion

This paper proposed a novel *multivariate* stability assessment technique consisting of (i) identifying a PCA model and control limits on reference data reflecting stable sludge and (ii) monitoring the start-up of subsequent experiments using control charts. The proposed approach only requires a rough variable selection by removing variables which are known to be unrelated to sludge stability. The stability information in the remaining variables is summarized in one scalar stability index (the SPE statistic), for which only one threshold needs to be defined. The proposed procedure was validated on the experimental data obtained by Van den Broeck et al. [1]. The control charts exhibit clear discrimination between the unstable and stable episodes.

Acknowledgements. Work supported in part by Project PFV/10/002 (OPTEC Optimization in Engineering Center) of the Research Council of the KU Leuven, Projects KP/09/005 (SCORES4CHEM) and KP/10/006 (AOPtimise) of the Industrial Research Council of the KU Leuven, and the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian Federal Science Policy Office. P. Van den Kerkhof is funded by a Ph.D grant of the agency for Innovation by Science and Technology (IWT). R. Van den Broeck is supported by the postdoctoral grant PDMK/11/108 from the KU Leuven Research Fund. J. Van Impe holds the chair Safety Engineering sponsored by the Belgian chemistry and life sciences federation essenscia. The authors assume scientific responsibility.

1. Van den Broeck, R., Van Impe, J., Smets, I.. Assessment of activated sludge stability in lab-scale experiments. *Journal of Biotechnology* 2009; **141**:147–154.

2. Barbusiński, K.. Adaptation of activated sludge to laboratory research conditions. *Environment Protection Engineering* 1991;**17**:57–65.

3. Barbusiński, K., Miksch, K.. Relationship between organic loading and some properties of activated sludge. *Journal of Chemical Technology and Biotechnology* 1997;**69**:357–361.

4. Higgins, M., Novak, J.. Characterization of exocellular protein and its role in bioflocculation. *Journal of Environmental Engineering* 1997; **123**(5):479–485.

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016

5. Jollife, I.. *Principal Component Analysis*. Springer Verlag; 1986.

6. Jackson, J.. *A user's guide to Principal Component Analysis*. John Wiley & Sons; 1991.

7. Lee, D., Vanrolleghem, P.. Adaptive consensus principal component analysis for on-line batch process monitoring. *Environmental Monitoring and Assessment* 2004;**92**:119–135.

8. Costa, J., Alves, M., Ferreira, E.. Principal component analysis and quantitative image analysis to predict effects of toxics in anaerobic granular sludge. *Bioresource Technology* 2009;**100**(3):1180–1185.

9. Mesquita, D., Amaral, A., Ferreira, E.. Characterization of activated sludge abnormalities by image analysis and chemometric techniques. *Analytica Chimica Acta* 2011;**705**:235–242.

10. Van den Broeck, R., Krzeminski, P., Van Dierdonck, J., Gins, G., Lousada-Ferreira, M., Van Impe, J., et al. Activated sludge characteristics affecting sludge filterability in municipal and industrial MBRs: Unraveling correlations using multi-component regression analysis. *Journal of Menbrance Science* 2011;**378**:330–338.

11. Li, B., Morris, J., Martin, E.. Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* 2002;**64**(1):79–89.

12. Nomikos, P., MacGregor, J.. Multivariate SPC charts for monitoring batch processes. *Technometrics* 1995;**37**(1):41–59.

13. Ramaker, H., van Sprang, E., Westerhuis, J., Smilde, A.. The effect of size of the training set and number of principal components on the false alarm rate in statistical process monitoring. *Chemometrics and Intelligent Laboratory Systems* 2004;**73**(2):181–187.

14. Arteaga, F., Ferrer, A.. Dealing with missing data in MSPC: several methods, different interpretations, some examples. *Journal of Chemometrics* 2002;**16**(8-10):408–418.

Journal homepage: http://www.journals.elsevier.com/process-biochemistry
Original file available at: http://dx.doi.org/10.1016/j.procbio.2013.07.016